Exploring the Dynamics of Transnational Financial Flows through Computational Linguistics

Forecasting Remittance Patterns among the Indian Diaspora

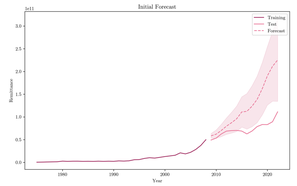
Arman Luthra | Faculty Sponsor: Nicholas Snow
Department of PPE, Wabash College

Introduction

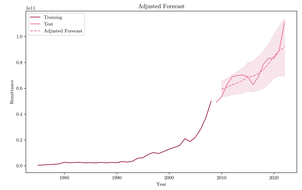
Remittances, significant financial inflows from migrant workers to developing countries, are vital to economies like India, the top global recipient. Accurately forecasting these remittances is essential for policy and economic planning, despite the complexity of prediction due to various influencing factors. In India, remittances substantially impact foreign exchange reserves, GDP, and economic growth, as well as household spending on essentials like food, education, and healthcare. The COVID-19 pandemic highlighted the vulnerability of these flows to external shocks. This research aims to improve remittance forecasting in India using advanced methods, combining natural language processing (NLP) with Seasonal Autoregressive Integrated Moving Average (SARIMA) time series modeling. SARIMA models, commonly used for forecasting, benefit from the integration of NLP's sentiment analysis, providing insights beyond structured data. By combining quantitative SARIMA forecasts with qualitative sentiment analysis, this interdisciplinary study seeks to enhance prediction accuracy, aiding policymakers in managing remittances, crucial for India's economy and its role as a remittance leader.

Results

The undertaken research is geared towards exploring the potential benefits of integrating Natural Language Processing (NLP) techniques with Seasonal Autoregressive Integrated Moving Average (SARIMA) models. The objective was to improve the accuracy of predicting transnational financial flows, in this case, remittances to India. The data used encompassed historical remittance data, and news articles related to the Indian diaspora and remittances. The initial phase of the research saw the application of a standard SARIMA (1,2,1)(1,1,1,12) model to the remittance data. However, the model's forecasts yielded a Mean Absolute Percentage Error (MAPE) of about 62.96%.

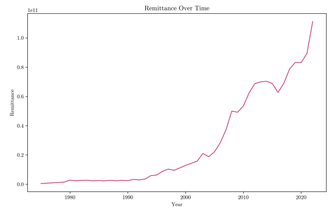


The research then ventured into the novel approach of integrating NLP driven sentiment analysis with the SARIMA model. The sentiment scores derived from the news articles acted as an adjustment factor in the forecasting model. This innovative methodology yielded impressive results, with the MAPE falling dramatically to approximately 6.25%, highlighting a significant improvement in the forecast's accuracy.



Methods

Stationarity Test



Before fitting the ARIMA model to this data, it was crucial to ensure its Stationarity. A stationary process has the property that the mean, variance, and autocorrelation structure do not change over time. Stationarity can be tested using the Augmented Dickey-Fuller (ADF) test.

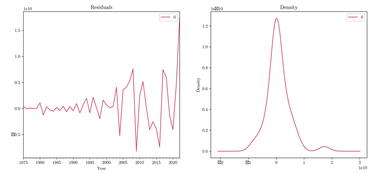
$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \delta_1 \Delta Y_{t-1} + \delta_2 \Delta Y_{t-2} + \cdots + \delta_p \Delta Y_{t-p} + \varepsilon_t$$

For the current data, since the initial ADF tests resulted in a p-value greater than 0.05, first-order differencing was applied. After applying first-order differencing, the ADF test was performed again and still, the p-value was greater than 0.05. This indicated that the series remained non-stationary, necessitating the application of second-order differencing. Subsequently, the ADF test confirmed the stationarity of the series as the p-value became significantly less than 0.05.

Fitting the ARIMA model

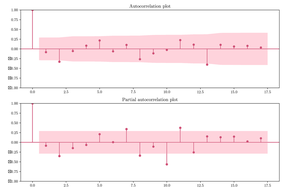
After ensuring the stationarity of the series, an Autoregressive Integrated Moving Average (ARIMA) model was fitted to the data. ARIMA is a popular time series forecasting model used to predict future values based on past observations. The ARIMA model combines three components: autoregression (AR), differencing (I), and moving average (MA). It is mathematically represented as:

$$(1 - \sum_{i=1}^{p} \varphi_i L^i)(1 - L)^d Y_t = \mu + (1 + \sum_{i=1}^{q} \theta_i L^i)\varepsilon_t$$



Incorporating Seasonality

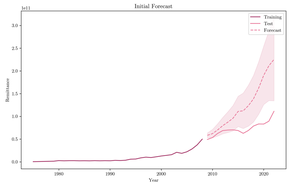
The PACF plot suggested that the first lag is significant, hence p = 1. From the ACF plot, it is observed that the first lag is significant, hence q = 1. Therefore, the ARIMA model with parameters (1,2,1) was fit to the data.



To further enhance the accuracy of the model, it was decided to utilize the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The SARIMA model is an extension of the ARIMA model that explicitly supports univariate time series data with a seasonal component.

$$(1 - \sum_{i=1}^{p} \varphi_i L^i)\,(1 - L)^d\,(1 - \sum_{i=1}^{P} \Phi_i L^{is})\,(1 - L^s)^D\, Y_t = \mu + (1 + \sum_{i=1}^{q} \theta_i L^i)(1 + \sum_{i=1}^{Q} \theta_i L^{is})\,\varepsilon_t$$

Upon careful examination of the graph, it became apparent that the forecasted remittances for the test period demonstrate a loose alignment with the actual remittances. This observation implied that the SARIMA model's predictive performance may be suboptimal for this period, possibly due to the prevailing economic conditions following the 2008 crash.



NLP Integration

This study collected a dataset of 1000 articles published in 2008, denoted as D. Each article is represented as $d_i \in D$, where i is the index of the article. Before performing sentiment analysis, the text data was preprocessed to make it suitable for analysis. The preprocessing steps involved tokenization, stop words removal, and lemmatization. The preprocessed version of an article d_i is denoted as $p(d_i)$.

Sentiment Analysis

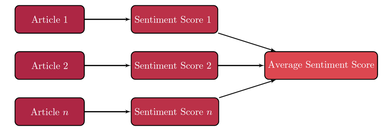
This study applied sentiment analysis to each preprocessed article $p(d_i)$. The sentiment analysis model assigned a sentiment score to each article, indicating its positive, negative, or neutral sentiment. The sentiment score of article d_i is denoted as $S(d_i)$.

Aggregating the results

After obtaining sentiment scores for all the articles, the results were aggregated to compute an average sentiment score for the entire dataset. The average sentiment score A was calculated using the following formula:

$$A = 1/N \sum_{i=1}^{N} S(d_i)$$

where N is the total number of articles (in this case, N = 1000).



Conclusion

The stark contrast in the performance of the two models underscores the potential advantages of leveraging NLP techniques in time series forecasting models. The findings of this study make a strong case for the integration of qualitative factors such as sentiment scores in remittance forecasting. This research opens new avenues in combining computational linguistics with econometrics,
promising to refine forecasting models and aid in creating more precise financial projections. Moreover, the successful integration of sentiment scores paves the way for further exploration in this direction. Other qualitative factors can also be considered for future research, expanding the scope of predictive models to encapsulate the nuances and intricacies of financial markets better. In conclusion, this research's findings point towards an exciting intersection of data science and financial forecasting. The significant reduction in forecasting error through the integration of sentiment scores demonstrates the transformative potential of NLP in enhancing financial forecasting models, making them more robust and accurate.

References

1. Budiharto, W. (2021). Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM). Journal of Big Data, 8(1), 47. https://doi.org/10.1186/s40537-021-00430-0
2. Guha, P., Islam, B., & Hussain, M. A. (2021). COVID-19 lockdown and penalty of joblessness on income and remittances: A study of inter-state migrant labourers from Assam, India. Journal of Public Affairs, 21(4), e2470. https://doi.org/10.1002/pa.2470
3. Isoto, R. E., & Kraybill, D. S. (2017). Remittances and household nutrition: evidence from rural Kilimanjaro in Tanzania. Food Security, 9(2), 239–253. https://doi.org/10.1007/s12571-017-0656-4
4. Ivanov, A., & Gocheva-Ilieva, S. (2013). Short-time particulate matter PM10 forecasts using Predictive Modeling Techniques. AIP Conference Proceedings. https://doi.org/10.1063/1.4827230
5. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1–35. https://doi.org/10.1145/3560815
6. Lu, J., & Meyer, S. (2020). Forecasting flu activity in the United States: Benchmarking an endemic-epidemic beta model. International Journal of Environmental Research and Public Health, 17(4), 1381. https://doi.org/10.3390/ijerph17041381
7. Lucas, R. E., & Stark, O. (1985). Motivations to remit: Evidence from Botswana. Journal of Political Economy, 93(5), 901-918.
8. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CORENLP Natural Language Processing Toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. https://doi.org/10.3115/v1/p14-5010
9. Mohapatra, S., & Ratha, D. (2010). Forecasting migrant remittances during the global financial crisis. The World Bank. https://doi.org/10.1596/1813-9450-5512
10. Personal remittances, received (current US\$) - India. World Bank Open Data. (n.d.). Retrieved from https://data.worldbank.org/indicator/BX.TRF.PWKR.CD.DT?locations=IN